2020 embedded VISION summit®

Cadence Tensilica Edge AI Processor IP Solutions for Broad Market Use Cases

Pulin Desai
Group Director, Vision & AI Product Marketing
September 2020

cadence®

Cadence Tensilica Processor and DSP IP Business

2020 embedded VISION summit

TENSILICA® CUSTOMERS

7B+ Processors SHIPPING Annually

DSP LICENSING REVENUE

#1 DSP IP LICENSING REVENUE

Processor LICENSING REVENUE

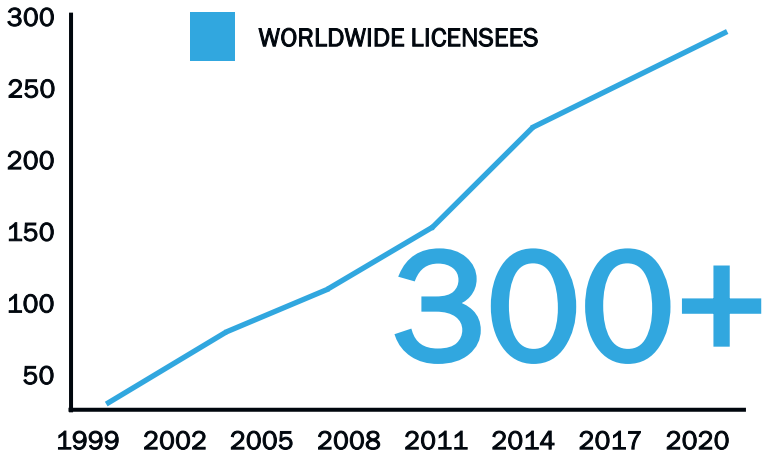
#2 Processor IP LICENSING REVENUE

SEMICONDUCTORS

19

19 of the Top 20 SEMICONDUCTOR **VENDORS** USE TENSILICA

TENSILICA LICENSEES



GLOBAL ECOSYSTEM

200+ ECOSYSTEM PARTNERS

cadence®

Examples of Edge AI: Voice & Vision

2020 embedded VISION summit

Voice AI

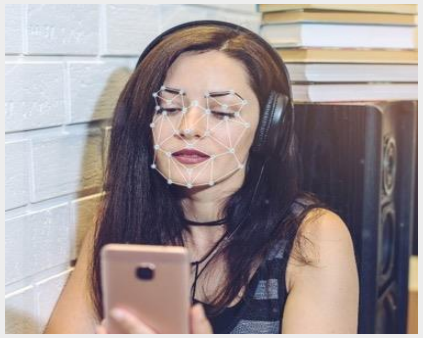
Smart Watch

Mobile

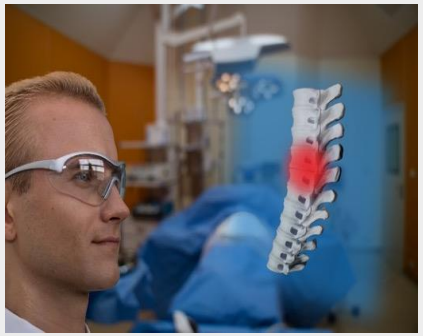
Smart Speaker

Headphones/Hearables

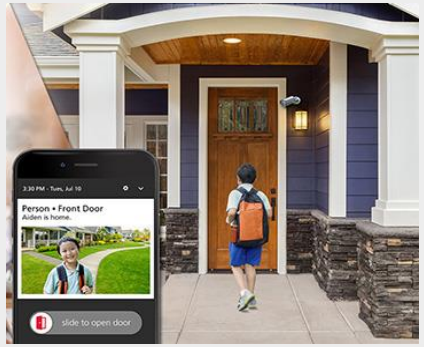
Vision AI



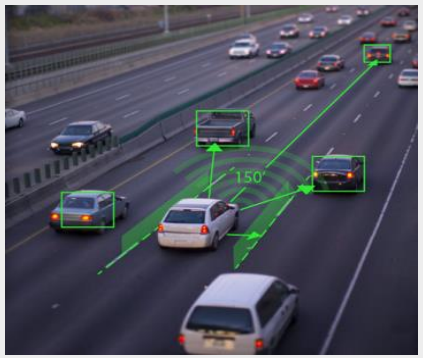
Mobile



AR/VR



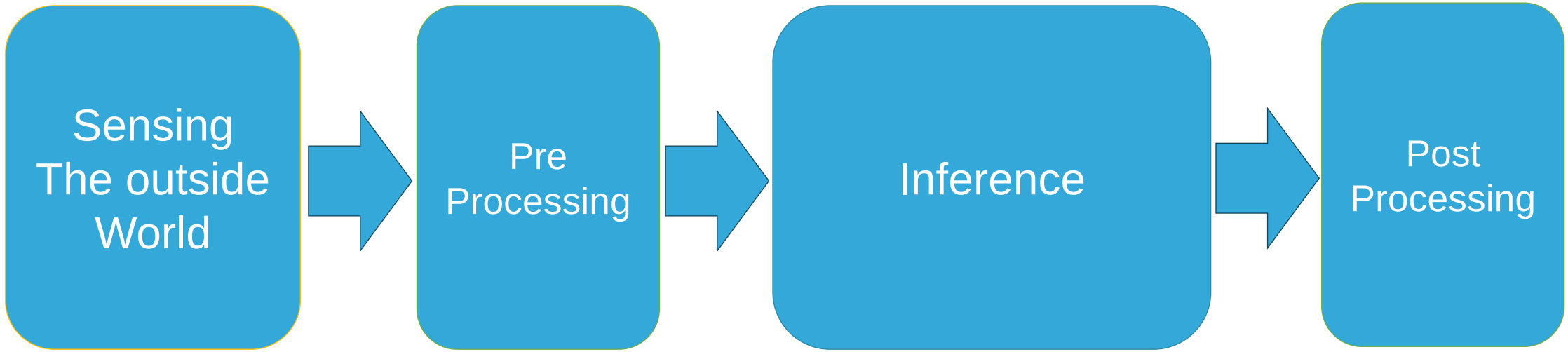
Smart Surveillance



Autonomous Vehicles

cadence®

Processing Flow for Inference in the Embedded System



- Where does data come from?
  - Image sensor
  - Radar
  - Lidar
  - Microphone

  } Vision

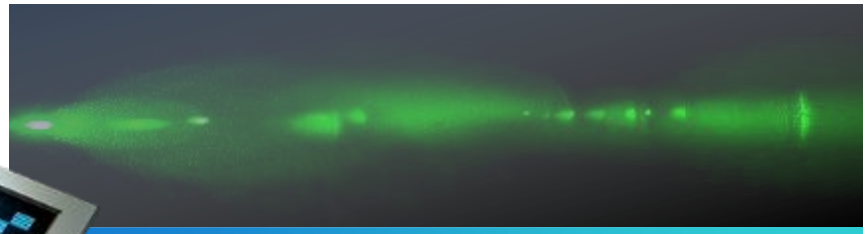
  Voice
- Preparing the data for Analysis
- AI to analyze the incoming data
- AI has other layers which are beyond convolution
- Any post processing of data

- Any sensing application always has a pre & post processing
- Need to solve the whole data path

Machine Learning Innovations Using Microphone as a Sensor

Voice UI

- Voice detection
- Voice trigger/ voice command
- Voice recognition
- Noise reduction
- Speech enhancement
- Speaker ID

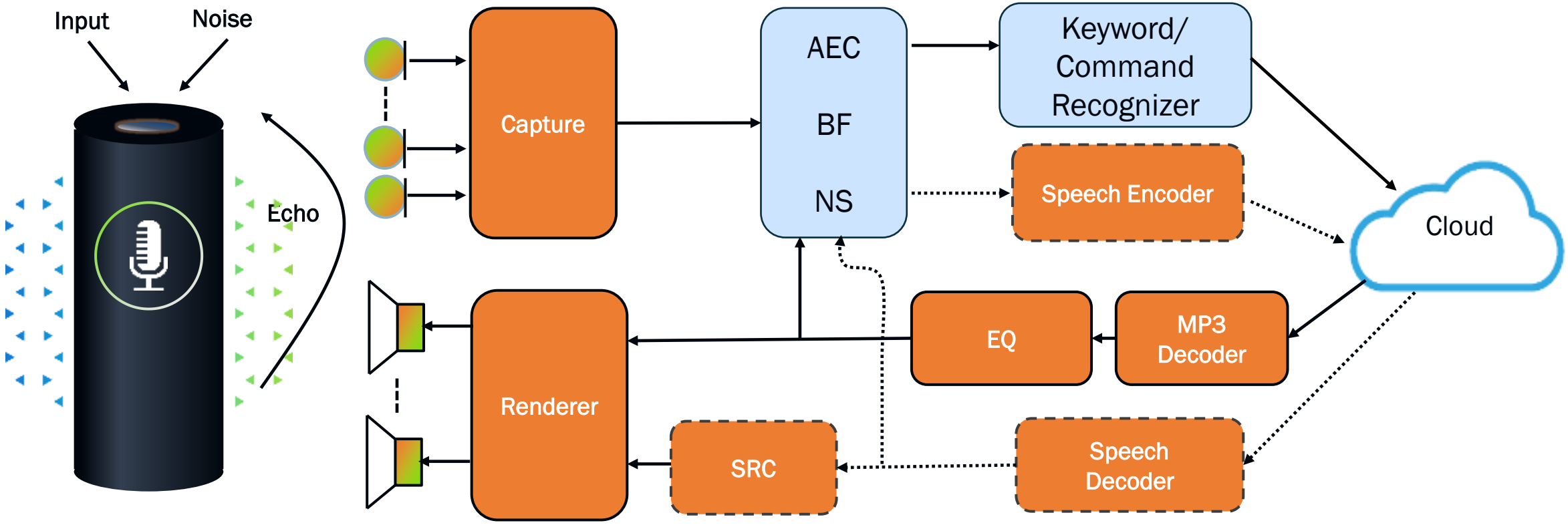


Sound Analytics

- Environment
- Sound classification
- Audio scene detection
- Baby cry
- Machine failure detection
- Song ID

Privacy, latency, power, and availability of network drives voice UI and other ML applications to edge devices

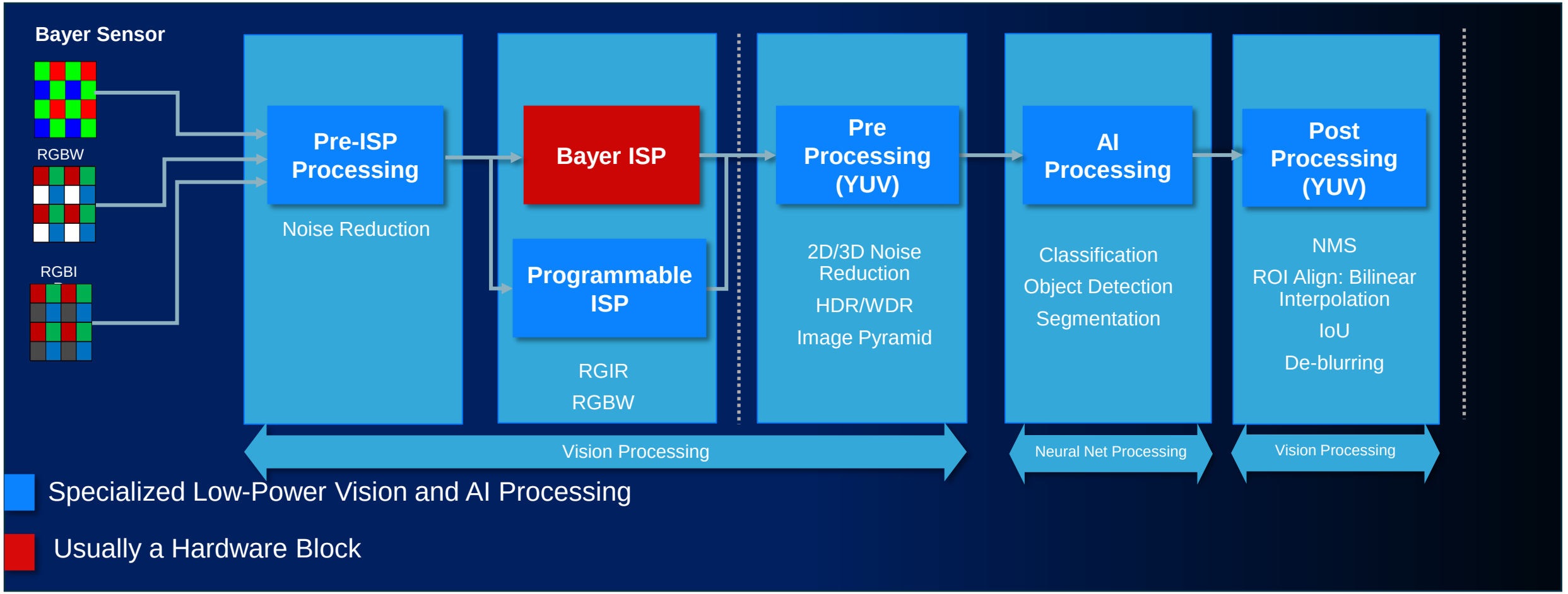
System Integration – Smart Speaker Example



AI Workload
Keyword /command recognizer
Voice enhancement & Noise suppression

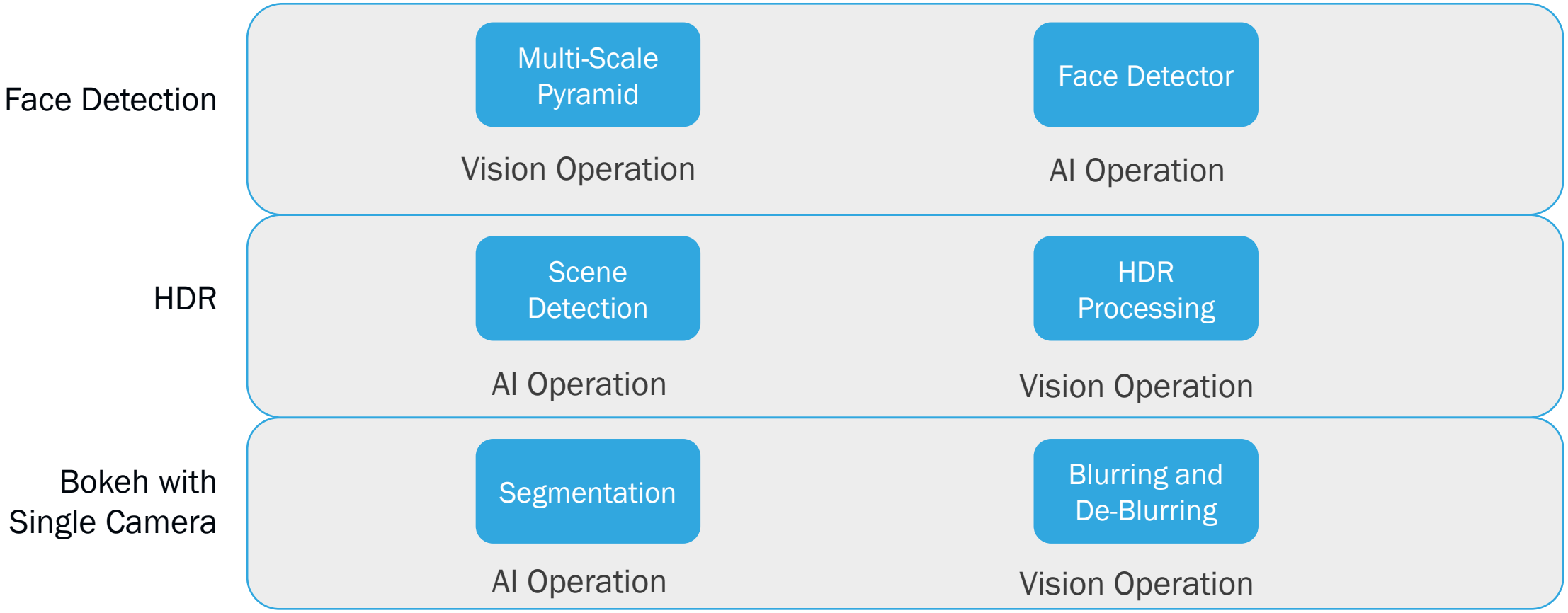
Non-AI Work load but needs a DSP
EQ, MP3 Decoder, Speech decoder are
Classical audio algorithms

Computer Vision and Image Processing Pipeline



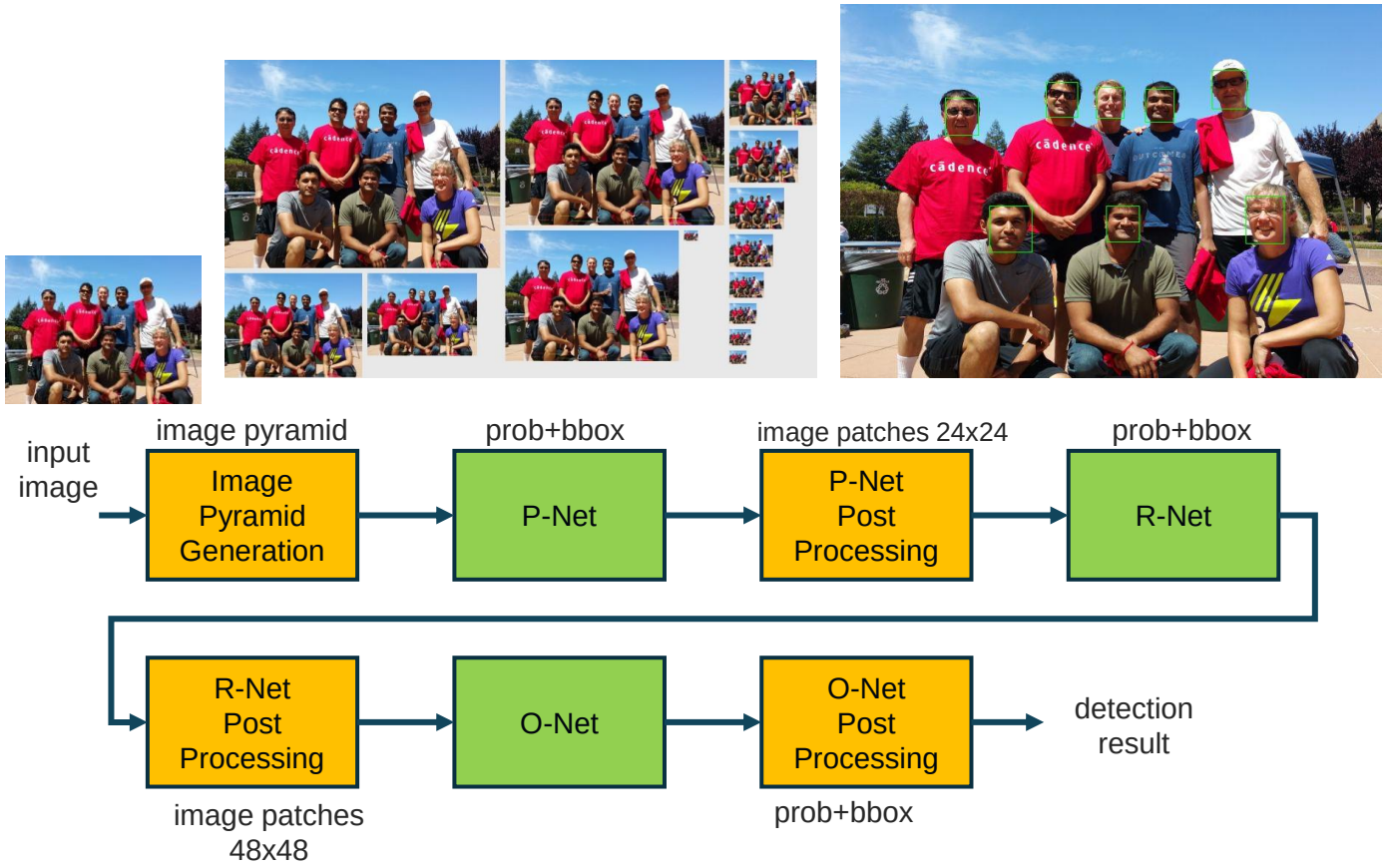
- Any Vision application has a pre & post processing

Vision Applications: Mix of Vision and AI



- All use cases still have mix of vision and AI operations
- Need for both vision and AI processing in the camera pipeline

Multi Net Application Case Study
MTCNN: A Face Detection Neural Network

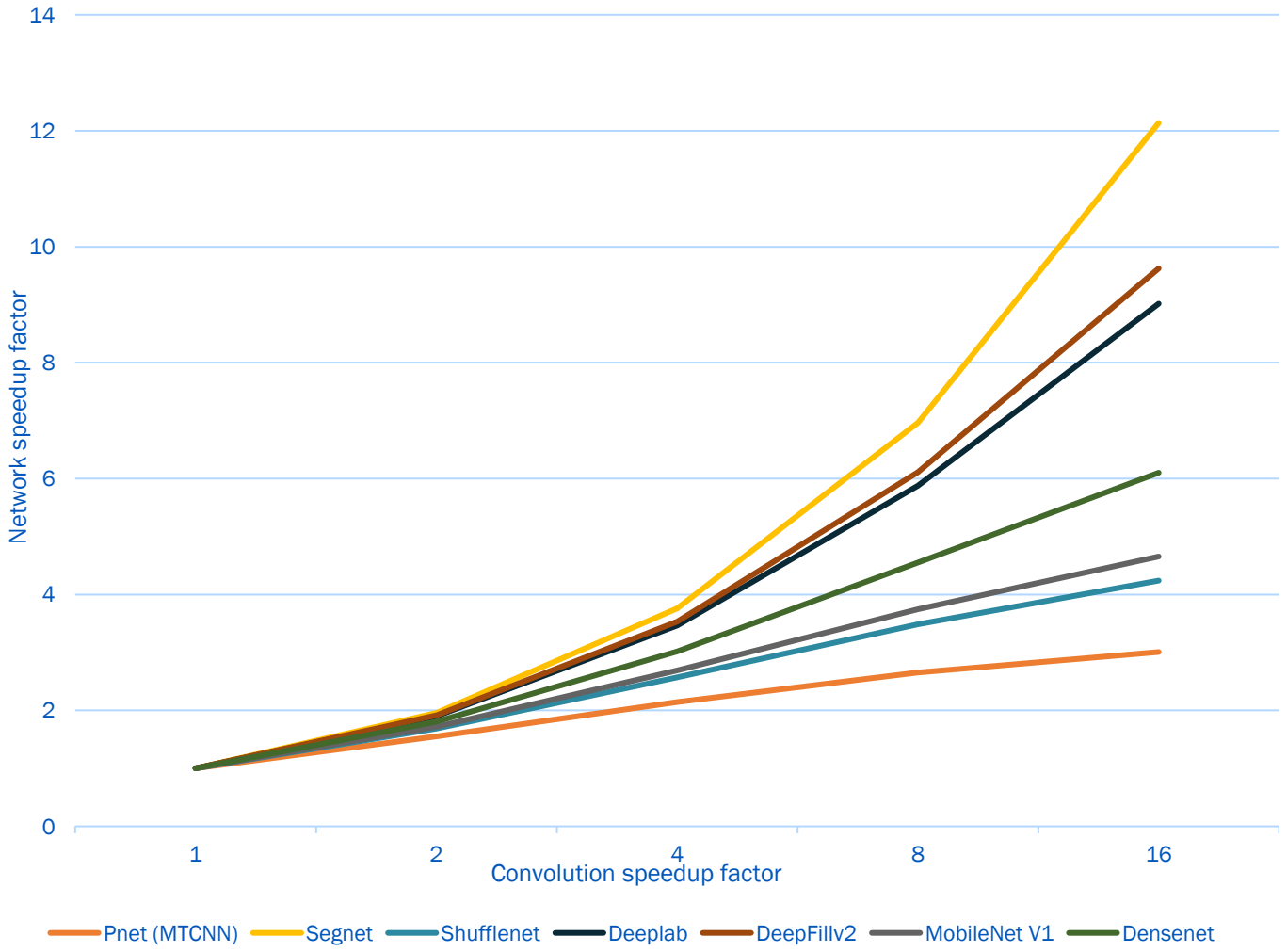


In MTCNN use case with 3 separate nets (P/O/R), glue processing is inserted between the nets to perform end to end face detection

Even if entire networks are sped up the application speedup will be bottlenecked by glue processing which does not need to be "standardized"

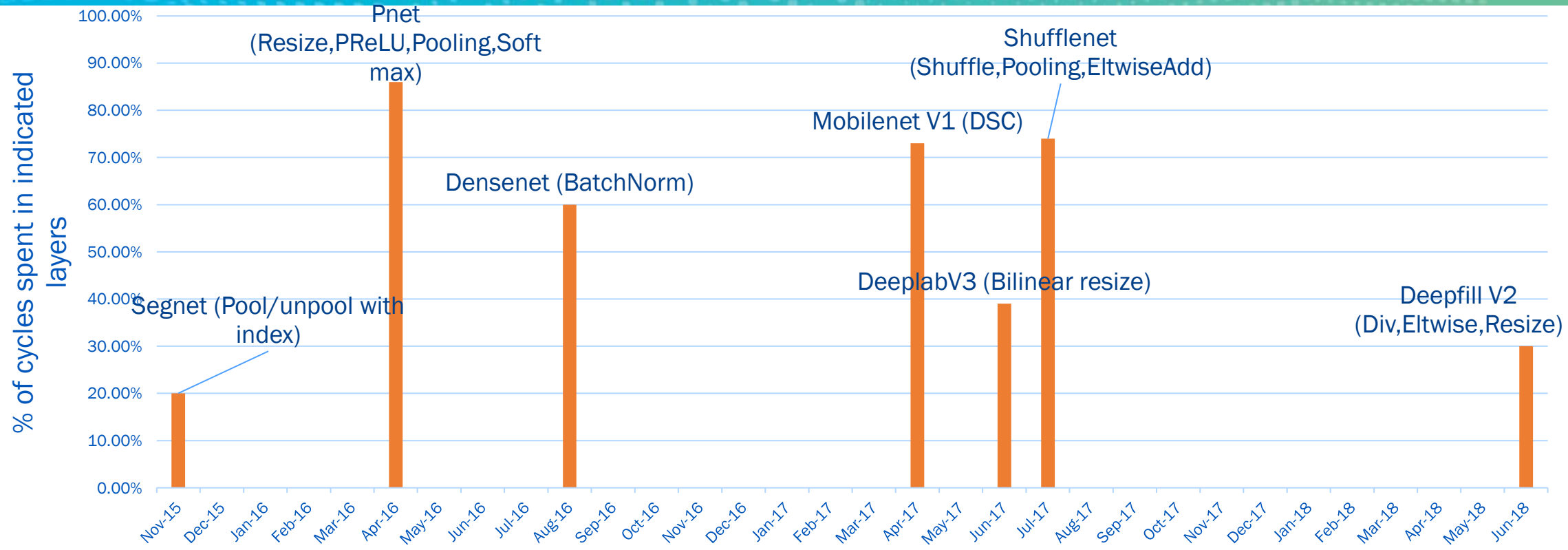
A high performance DSP targeted at Vision applications can keep up with the evolving glue processing requirements

Hypothetical Speedup when only Convolution is Accelerated



- Convolutions are common target for acceleration with hardware
- As convolutions speed up, different networks will achieve varying amounts of overall speedup
  - Some networks achieve only modest speedups bottlenecked by "other" layers
- "Other" layers need to keep up with convolutions acceleration

Significant Cycle Consuming Non-Convolution Layers



- The set of “other” layers is large and continuously evolving
  - Non-convolution layers can be 20% to as high as >80% of the cycles
- A high performance DSP targeted at NN applications can keep up with the evolving requirements

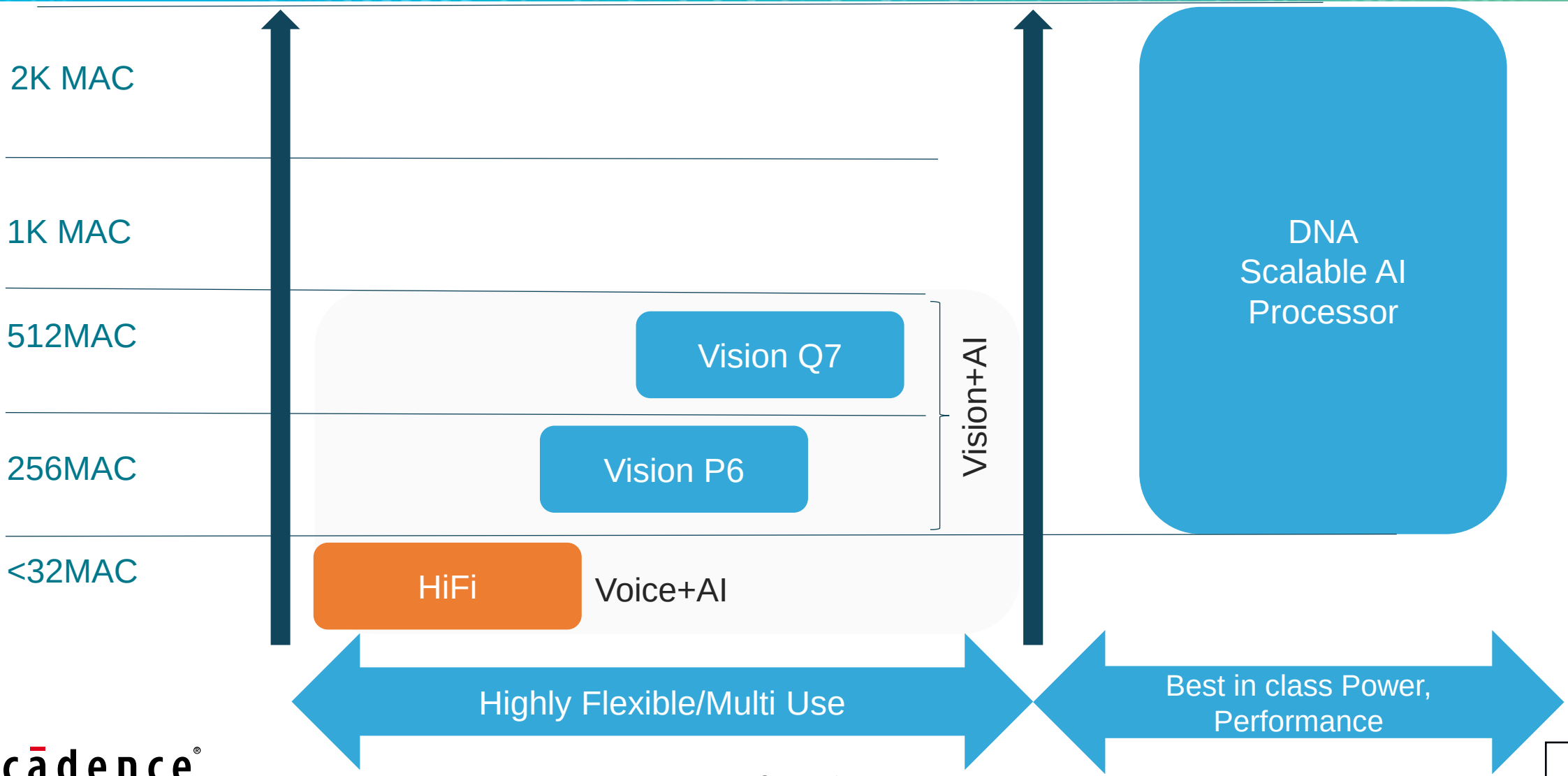
Neural Network Deployment Options in Embedded Systems

2020 embedded VISION summit

Metric: / Option	Flexibility	Power	Performance	Time to Market
CPU	✓ Highly flexible	Very high	Depends	Fast
CPU+GPU	✓ Highly flexible	Very high	Depends but could be high	Fast
Hardware Development	Not flexible	Low	High	Long development cycle and verification
Programmable DSP	✓ Highly flexible	✓ Low	Depends	✓ Fast
Programmable DSP+ HW accelerator	✓ Highly flexible	✓ Low	High	✓ Fast

cadence®

Tensilica AI Portfolio: AI Inference at the Edge



Voice AI
Vision AI

HiFi DSP Is Ideal for ML Solutions on the Edge

2020 embedded VISION summit

Tensilica® HiFi DSP is the leading IP solution for audio, speech, and machine learning applications

Cadence enables developers to train and seamlessly deploy their machine learning applications on HiFi DSPs using TensorFlow Lite for Microcontrollers

HiFi DSP has extensive production-ready software solutions from Cadence and the largest ecosystem with 160+ software partners

Tensilica® Xtensa Audio Framework (XAF) accelerates integration of ML and traditional audio workloads, reducing time to market

cadence®

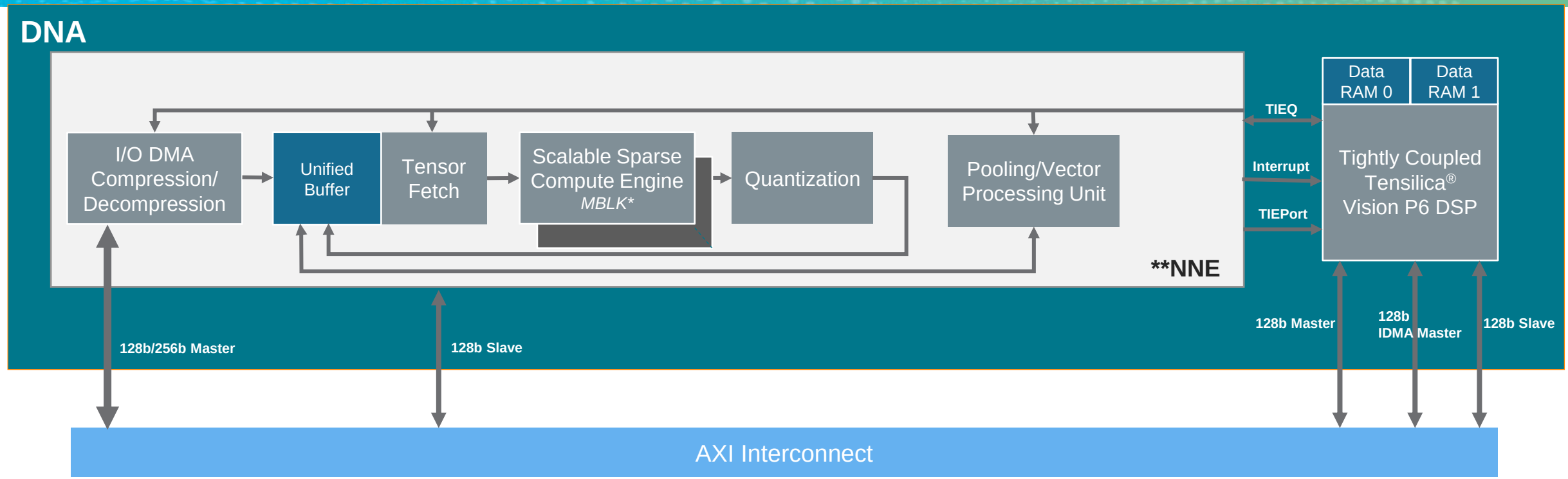
Vision P6/Q7: Ideal DSP for Vision + AI

2020 embedded VISION summit

Vision Acceleration	• 512-bit SIMD • 5 VLIW Slots • 8bit/16bit/32bit int, SP and HP FP Support • 0.7 to 1.7 TOPS
AI Acceleration	• 256/512 8-bit MAC • FP16 support
High Data Throughput	• 1024 bit memory I/F • Scatter Gather Technology • Multi-channel iDMA • 128/256 bit Axi
Lib & Tools	• OpenCV Base imaging/vision Lib, SLAM Lib • OpenCL, Halide, OpenVX Support • NN Compiler (ONNX & GLOW) & Android NN API Support

cadence®

Tensilica DNA Processor: Single Processor for Vision & AI Sparse Compute Engine & Scalable



DNA = XNNE + Vision P6

- ✓ XNNE (Xtensa Neural Network Engine)
  - ✓ Scalable Sparse Engine
  - ✓ Scales from 256MAC to 2048MAC
  - ✓ Unified Buffer

Vision P6:

- ✓ VLIW SIMD DSP
- ✓ For future proofing & expanding
- ✓ Pre & Post Processing
- ✓ CV workload

Typical Vision + AI Work: On Tensilica DNA Processor

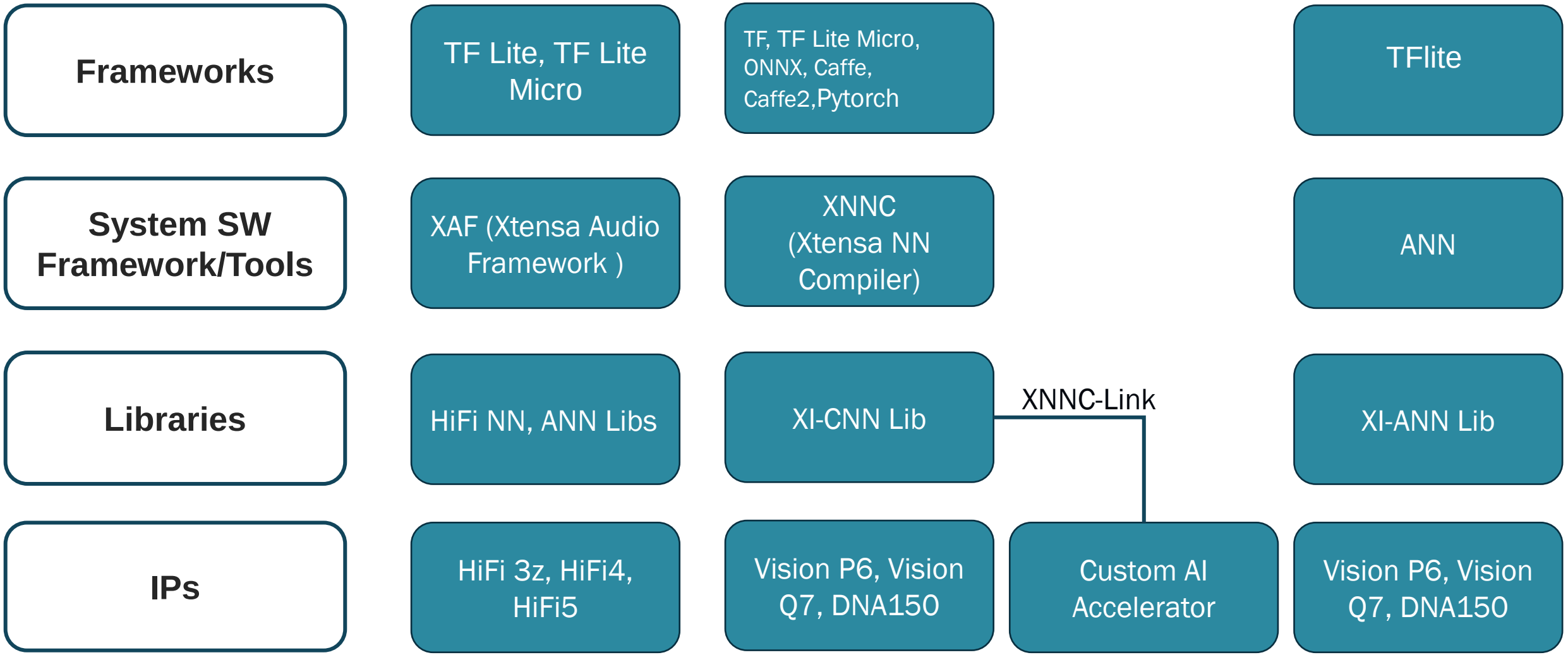
2020 embedded vision summit

Pre-Processing	AI Workload (Convolution)	Post-Processing
Runs on Vision P6	Runs XNNE HW of DNA Processor	Runs on Vision P6
Multi-scale Pyramid Color-space Noise Reduction	NN Workload Mostly various convolution Non-Convolution Layer	Non-Max Suppression (NMS) ROI Align: Bilinear Interpolation De-blurring HDR

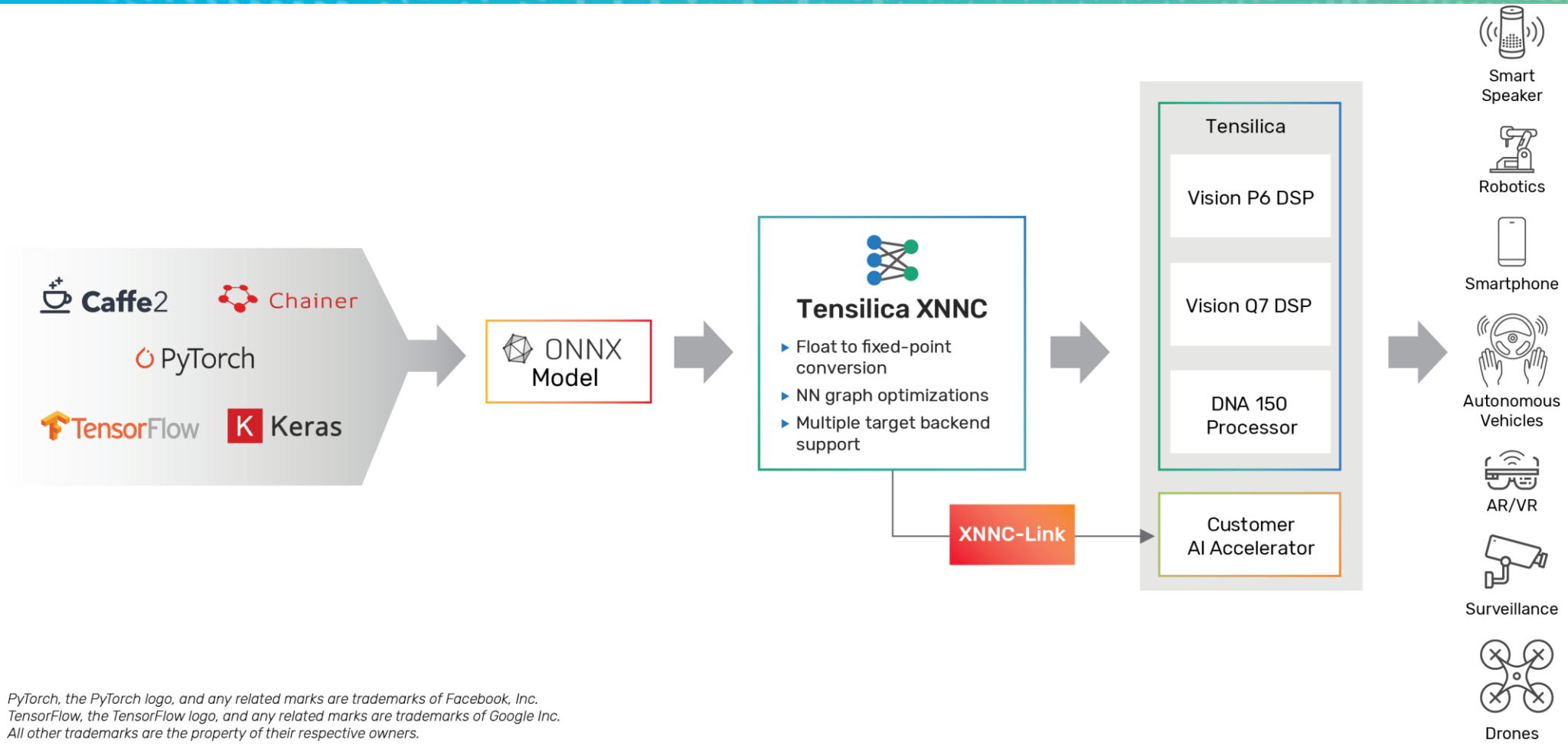
Example Networks:
Yolo, SSD, MaskRCNN, CRNN
Could be >10% cycles
Depends on # of detection
Depends on accuracy

cadence®

Tensilica AI SW Tools Portfolio



XNNC-Link
NN Code Generation for customer's AI Accelerator



Vision DSP + Customer's accelerator for your AI Solution

Edge-AI products with Cadence Tensilica AI IP

2020 embedded VISION summit

MEDIATEK

"MediaTek has confirmed that the P60 integrates a Cadence Vision P6 core for its AI accelerator." Source (1)

MEDIATEK

Mediatek i500 : " AI acceleration for APU based on **Tensilica Vision P6** (i500) " Source: (2)

TOSHIBA

"Toshiba Selects Cadence **Tensilica Vision P6** DSP as Image Recognition Processor for its Next-Generation ADAS Chip" Source (3)

Kneron

Kneron, the San Diego and Taipei-based low-power edge AI startup, "KL720 NPU IP integrated with **Cadence Tensilica Vision P6 DSP IP**" Source (4)

NXP

NXP i.MX-RT600
HiFi 4 for audio and voice processing

"NXP's enablement for Glow is tightly coupled with the Neural Network Library (NNLib) that Cadence provides for its **Tensilica HiFi 4 DSP** delivering 4.8GMACs of performance." Source: (5)

Baidu 百度

Baidu HongHu
HiFi 4 for audio and voice processing

"Baidu today released its new chipset Honghu at the annual AI Developer Conference in Beijing. It features **HiFi4** custom instruction set, dual-core DSP, and only 100mV power dissipation on average." Source: (6)

Source 1: https://www.eetimes.com/mobile-ai-race-unfolds-at-mwc/#
Source 2: https://www.cadence.com/en_US/home/company/newsroom/press-releases/pr/2019/toshiba-selects-cadence-tensilica-vision-p6-dsp-as-image-recogni.html
Source 3: https://www.mediatek.com/blog/mediateks-rich-iot-sdk-v20-0-release-available-now-for-i300-and-i500-chipset-series
Source 4: https://www.eetimes.com/kneron-raises-40m-for-next-gen-edge-ai-chip/
Source 5: https://media.nxp.com/news-releases/news-release-details/industrys-first-mcu-based-implementation-glow-neural-network
Source 6: https://en.pingwest.com/w/2549

cadence®

Summary: Cadence Tensilica Edge AI Processor IP Solutions

- Voice+AI & Vision+AI for edge requires pre & post processing in addition to neural network processing
- In addition to convolution layers (which are MAC heavy) other layers take considerable cycles
  - MAC acceleration only does not accelerate the AI performance
- Combination of hardware acceleration for convolution and programmable DSP provides the best solution for Edge AI workload
  - Single hardware can be used for both concurrent Vision+AI and Voice+AI workload
- Cadence Tensilica provides Vision+AI and Voice+AI HW IP with NN Compiler SW
  - Cadence Tensilica IP is shipping in large number of edge-ai products

Resource Slide

Cadence Resources

https://www.cadence.com/en_US/home/tools/ip/tensilica-processor-ip.html

https://ip.cadence.com/vision

https://ip.cadence.com/ai

External Resources

https://onnx.ai/

https://onnx.ai/supported-tools.html

https://www.edge-ai-vision.com/2019/10/cadence-demonstration-of-a-recurrent-neural-network-based-show-attend-and-tell-on-a-tensilica-dsp-platform/