Designing the Next Ultra-Low-Power Always-On Solution

Amol Borkar
Product Marketing Director, IPG TIP
Cadence Design Systems, Inc.

Cadence Tensilica IP: Business Overview and Recent Success

TENSILICA® CUSTOMERS

>40B

Processors

SHIPPED

DSP LICENSING REVENUE

#1 **DSP IP LICENSING REVENUE**

SEMICONDUCTORS

19

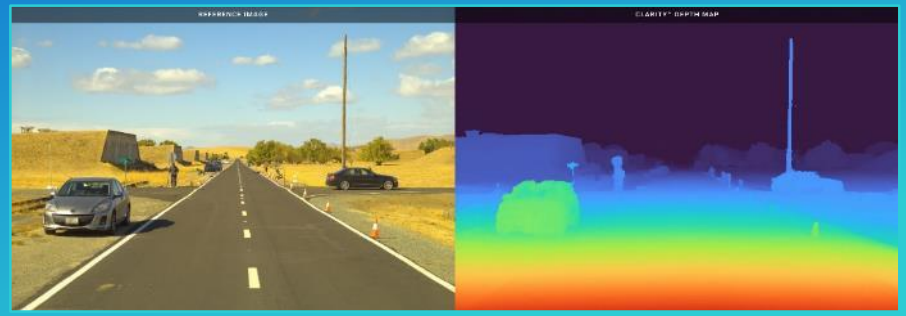
19 of the Top 20
SEMICONDUCTOR VENDORS
USE TENSILICA

GLOBAL ECOSYSTEM

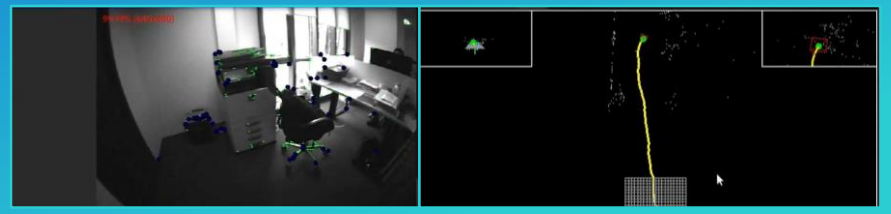
200+ **ECOSYSTEM PARTNERS**

UNISOC™

light



xvisio TECHNOLOGY



cadence

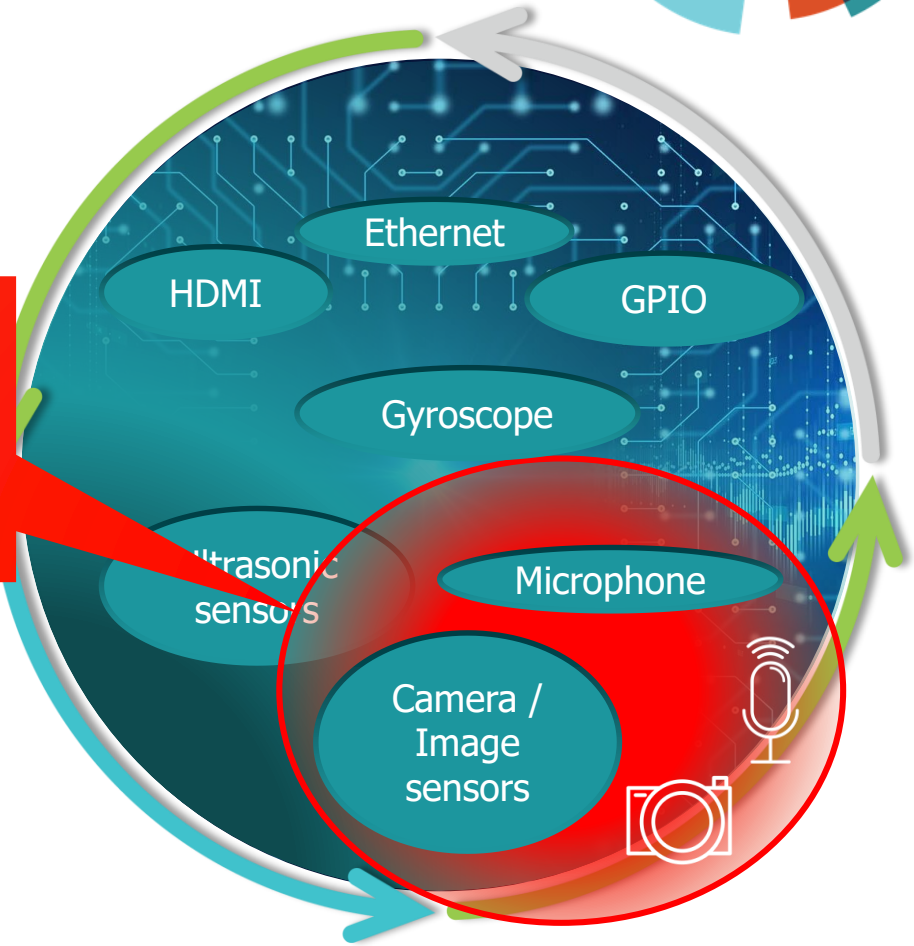
Processing and Sensor Trends

Automotive, robotics, drones, mobile, and AR/VR

Automotive	Robotics	Drone	Mobile	AR/VR
				
Different Types of Sensors: Image Sensors/Lidar/Radar			3D Sensors (TOF), Image Sensors	
Large Number of Image Sensors/High FPS/High Resolution				
Sensor Fusion				
3D Point Cloud/Capture				
NEW **Ultra-Low-Power Computer Vision/AI and Always-On Island**				
Computational Imaging Algorithms				

Always-On / Wake-Up Processing

- Protocols or methodologies to "wake up" the device has been around for a long time and is being used in many industries
- Various options for input sources: HDMI, GPIO, Gyroscope…
- Using **microphone** and **ultrasonic sensors** has been prominent in the past few years
  - Proliferated through AI for such usages with TFLm and TinyML
- Interest in using cameras for "visual wake words"
  - People detection, face detection, gestures, etc.
- Emphasis on vision and speech-based processing for embedded/IoT devices (some ultrasonic)



Always-On Primer

- Lightweight algorithms that constantly run or execute on device, typically as a precursor to more accurate (more compute-intensive) algorithms
- Also known as low power or standby mode or AON
- **Applications**


Device Unlock / Authenticate

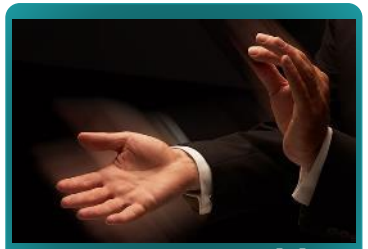

Video Doorbell

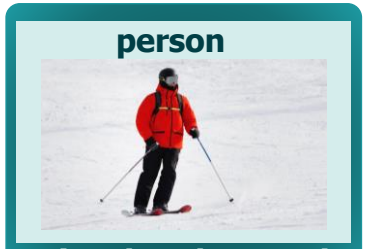

Gesture Control

IoT/Battery Powered Devices

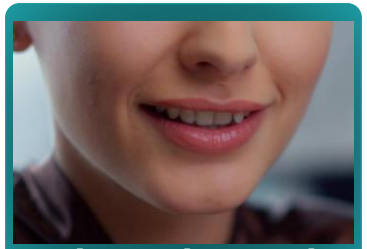
- **Ways to Accomplish This**


Face Detection


Gesture Recognition



Visual Wake Word


Voice Wake Words

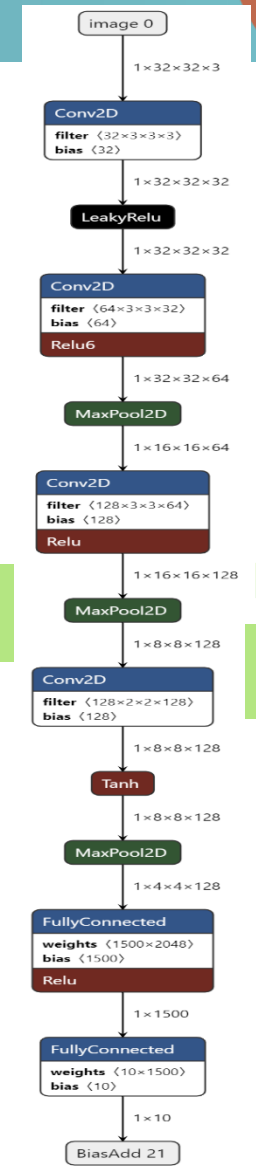

Fingerprint


AI

Always-On Requirements

- Basic functionality:
  - Continuously running or actively monitoring on the embedded device
  - Process inputs from cameras, microphones or other sensors in real-time (e.g., 5 – 20 fps)
  - Capable of handling traditional algorithms and lightweight neural networks
  - Ultra-low-power / energy consumption (e.g., <1 mW)
- Introduction of "lite" networks for always-on usages:
  - Designed to run on microcontroller class of devices (e.g., TensorFlow Lite for Micro)
- Common problems in always-on domain
  - Limited availability on public domain (not present in most common model zoos)
  - Expectation that model zoo networks have low energy consumption on always-on processors

Typical Networks	Total MAC count (MAC/Inf)
	Model Zoo Network
Face / Person / Object Detection	30 G
Speech / Voice	200 M
Gesture	10 G



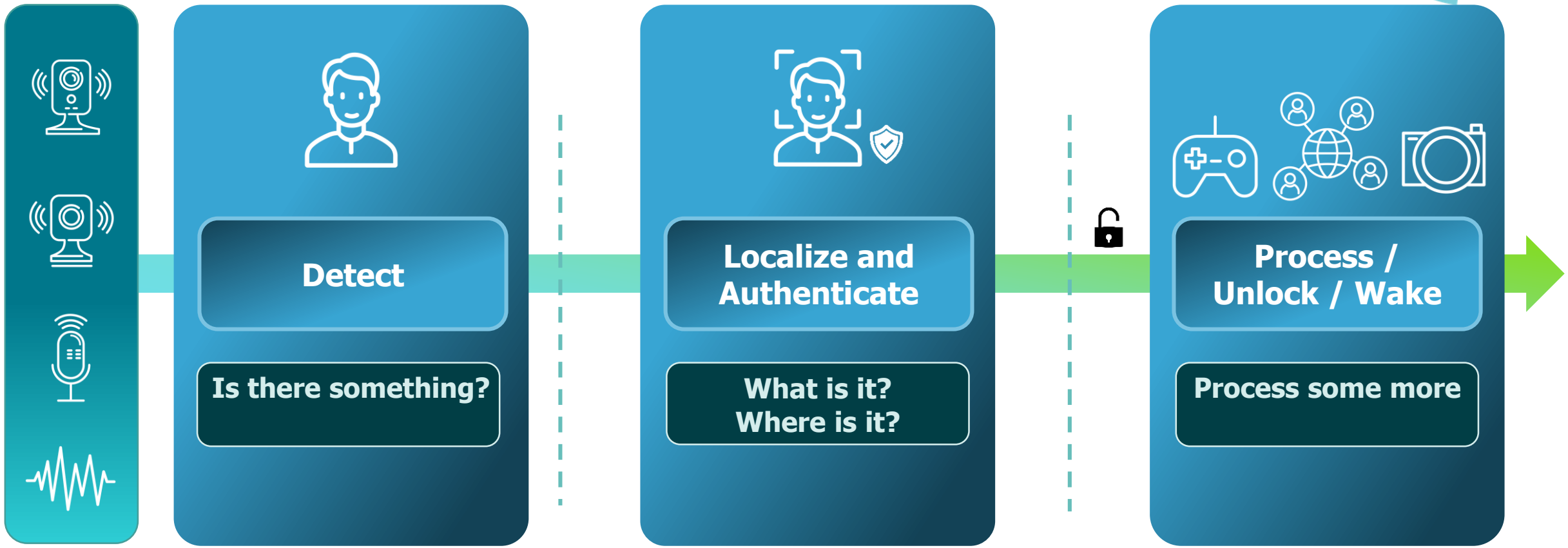
100 Layers

12 Layers

Networks for AON processors are **significantly** smaller than typical model zoo networks

**Approx values, does not depict any specific network(s)*
**Resnet50 vs TinyML Image Classification Networks*

3 Stages of Device Wake / Unlock



3 Stages of Device Wake / Unlock

Target segment for always-on	Detect	Localize and Authenticate	Process / Unlock / Wake
Sensor	Ultrasonic, Microphone, Camera, Fingerprint scanner	Microphone, Camera	Ultrasonic, Microphone, Camera, Touchscreen and more
AI Compute Requirements	< 250 GOPs	250 GOPs – 2 TOPs	> 2 TOPs
Processor	mCU or LP DSP	DSP, AI Engine, HWA	CPU, DSP, GPU, AI Engine, HWA
Key Features	Ultra low power / energy Area	Low power / energy Area	
Typical Applications	Keyword Spotting, Person/Face Detection, Visual Wake Words, Gestures	Automatic Speech Recognition, Person / Face recognition and identification, refined gesture processing	Pose / Gait recognition, 3D Face, Speech recognition, Gaming
SW Tools/Fwk	Bare-metal, TFLm	TensorFlow, TFLite, Pytorch, ONNX, Imaging, CV etc.	Too many to list

cadence

Focusing on Detection Stage

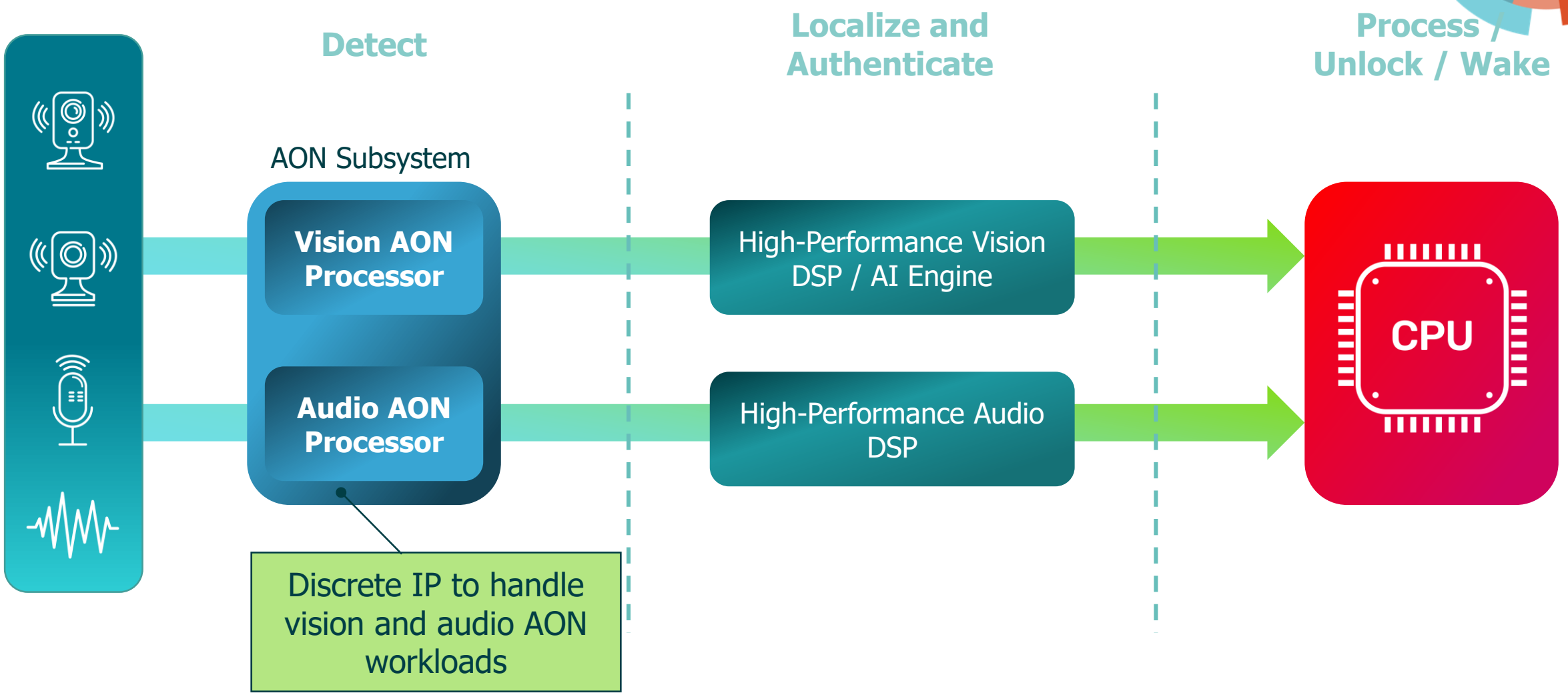
Lots of ongoing efforts

Emerging markets

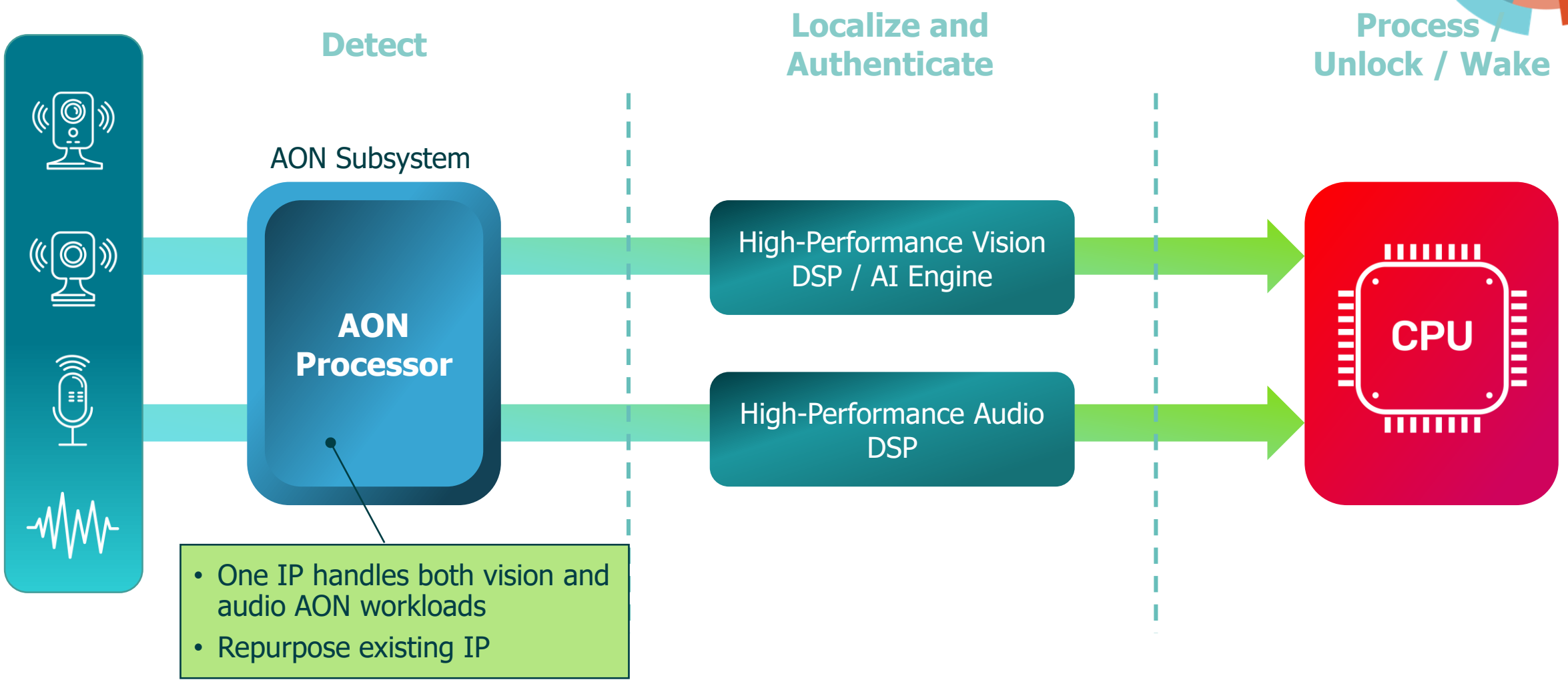
Market Classification	Entry-Level / Base	Mid-Range	Premium
Sensor(s)	Ultrasonic, Microphone, Low res. camera	Microphone or Camera or others	Microphone, Camera and others
Typical Applications	Proximity detection	Keyword Spotting, Person / Face detection …	Keyword Spotting, Person / Face Detection … **IP RE-PURPOSE**
Benefits	✓ Best Area, Power, Energy ✓ Simplified algorithms	✓ Leverage cutting edge AI algorithms ✓ Higher resolution for identification	✓ Area saving with IP repurpose ✓ Scalable and future-ready
Limitations	X Not enough res for user identify X Limited AI capabilities	X Able to run only one modality (Keyword Spotting, or Face Detection…)	X Increased power consumption

cadence

Architectural Considerations: Discrete Data Paths



Architectural Considerations: Combined Data Paths



What Have We Learned So Far? (AKA Requirements)

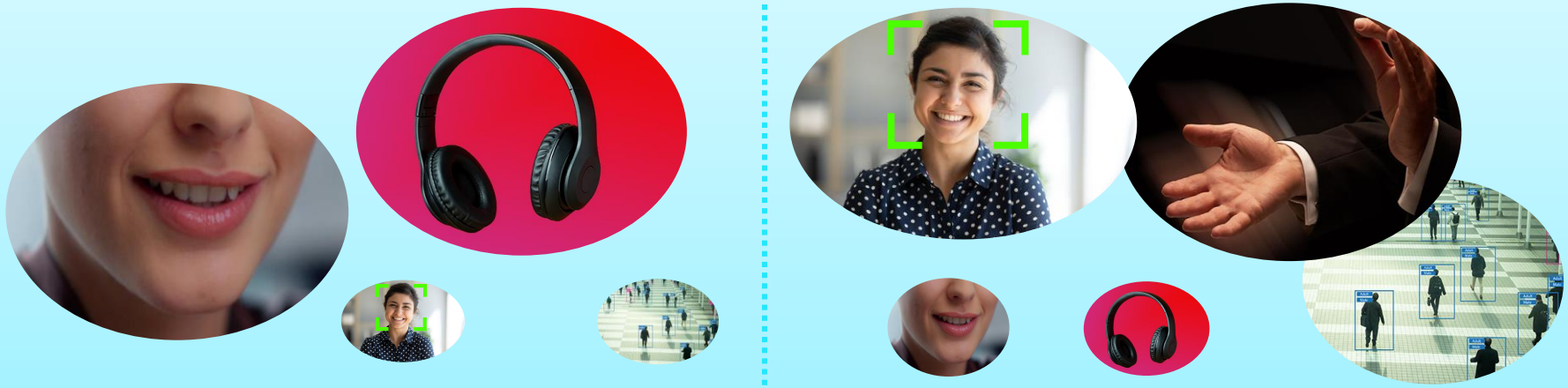
- There is a need to do always-on processing (AON)
- Area-efficient processor with ultra-low power consumption
  - Operate with extremely small memory footprint i.e., no DDR
- Trending towards discrete or single IP to do multi-modal processing
- Processing inputs from camera, microphone, ultra-sonic sensor, etc.
- Capable of computer vision, AI, speech, sound, FFTs, etc.
- Fortunately, Tensilica® has an answer …

Selection-Based on Dominant Processing

Sensors

Applications

Dominant Processing



Processor Choice

Audio / Speech Dominant Processing → Tensilica® HiFi DSP

Image / Camera Dominant Processing → Tensilica® Vision DSP

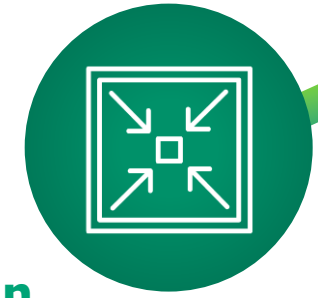
cadence

Leading Solution for Low-Energy Always-On

Scalability
Based on very successful Vision DSP architecture
N-way programming model

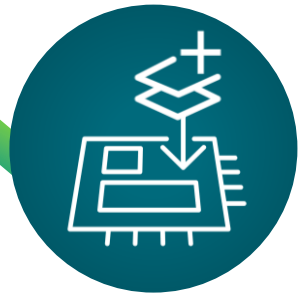
Tensilica® Vision P1 DSP
Best-In-Class Low-Energy DSP for Vision and AI DSP

Enables Compelling Use Cases
Always-on processing
Face detection
Smart image sensor
Foveated rendering
Proximity detection
Keyword spotting



Efficient Design
1/3 area and power of Vision P6 DSP

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



Scalable, Configurable, and Extensible Architecture
128-bit SIMD
Up to 128 MAC performance
Supports TIE for custom instructions

cadence

Always-On Face Detection (1)

Good
(CPU)

- ✓ CPU **continuously** performs face detection and recognition algorithms
- ✓ CPU recognizes face and unlocks phone
- ✓ ***Example:*** *Phone with CPU only*
- X CV+NN algorithms on CPU are not very efficient

Detect, Recognize, Authenticate

CPU or App Processor

Always-On Face Detection (2)

Better
(CPU + Vision DSP)

- ✓ Vision DSP **continuously** performs face detection and recognition algorithms
- ✓ Vision DSP recognizes face and wakes up host CPU to unlock phone
- ✓ ***Example:*** *Phone with CPU + Vision DSP*
- X DSP is performing face detection when no face is present

Tensilica®
Vision DSP or NPU

Detect, Recognize, Authenticate

CPU or App Processor

Always-On Face Detection (3)

- Low Resolution
- ISP Bypass
- Bayer / Mono

Tensilica®
Vision P1 DSP

There is a face

Best
(CPU + Vision DSP + Vision P1)

Tensilica
Vision DSP or NPU

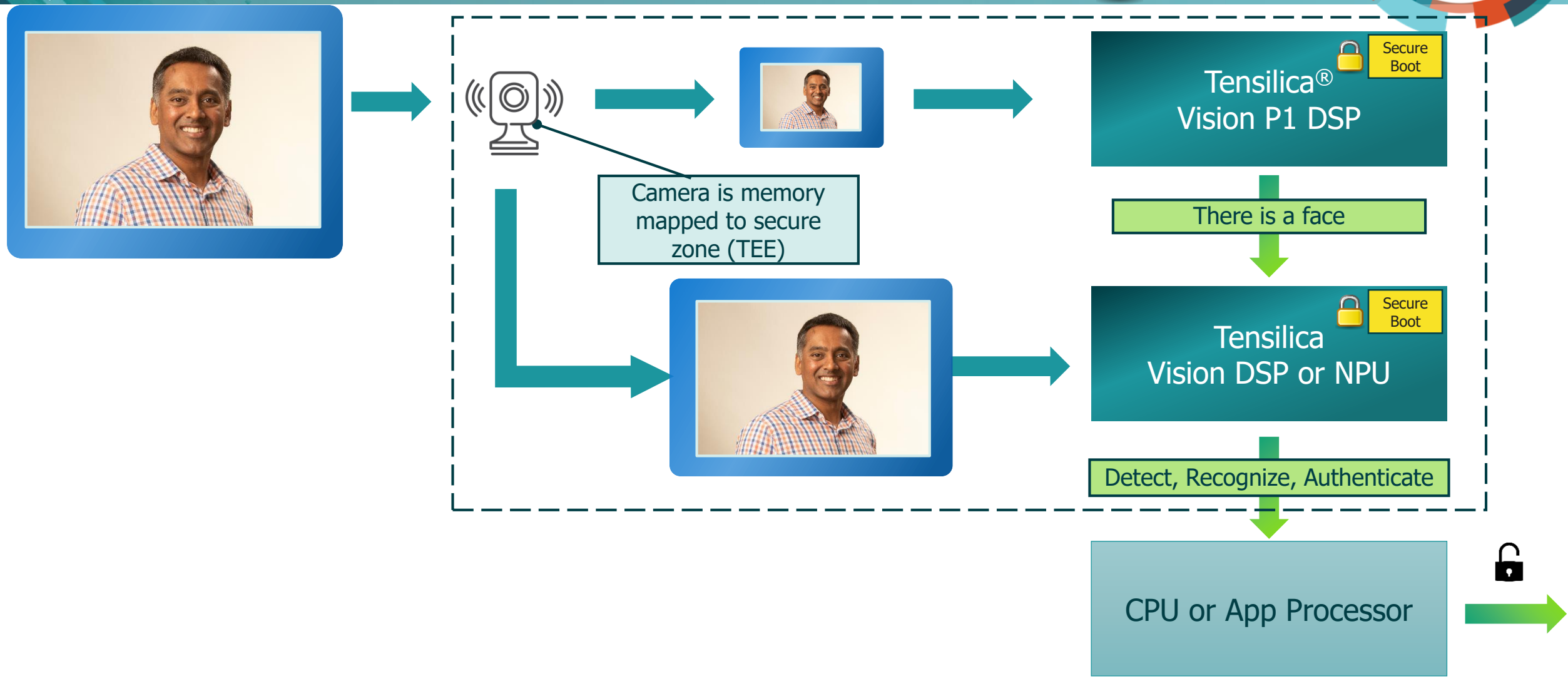
Detect, Recognize, Authenticate

CPU or App Processor

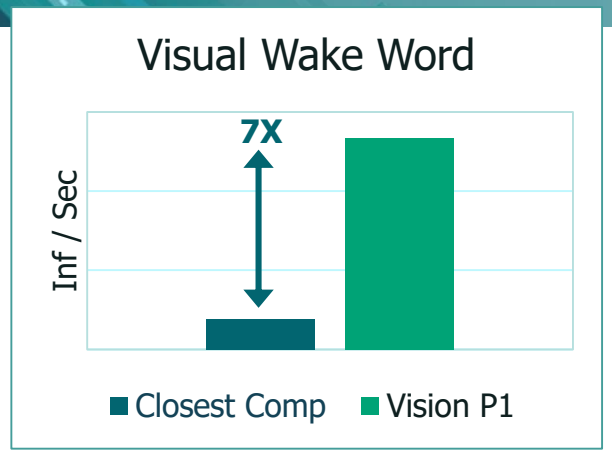
- ✓ Vision P1 DSP continuously performs **simplified** face detection algorithms
- ✓ Wakes up Vision DSP to do accurate face detection and recognition
- ✓ Vision DSP recognizes face and wakes up host CPU to unlock phone
- ✓ **Best energy efficiency: Vision DSP and CPU are only woken up as needed**

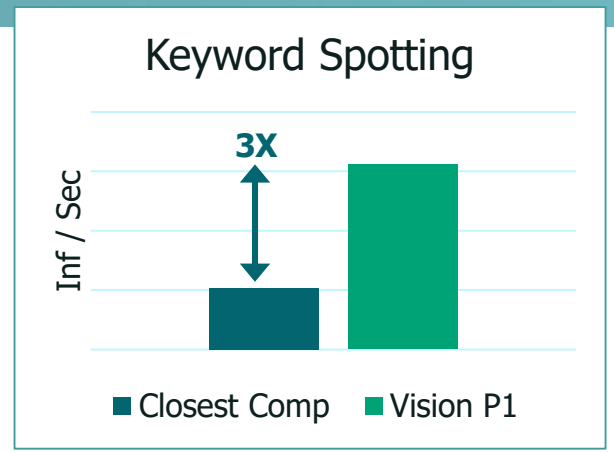
cadence

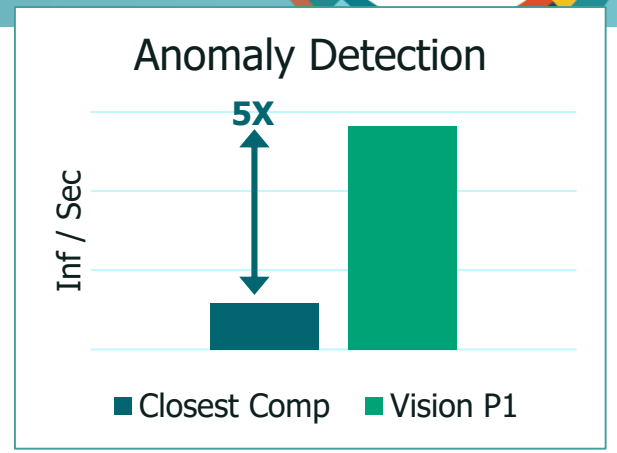
Always-On Face Detection with Secure Mode

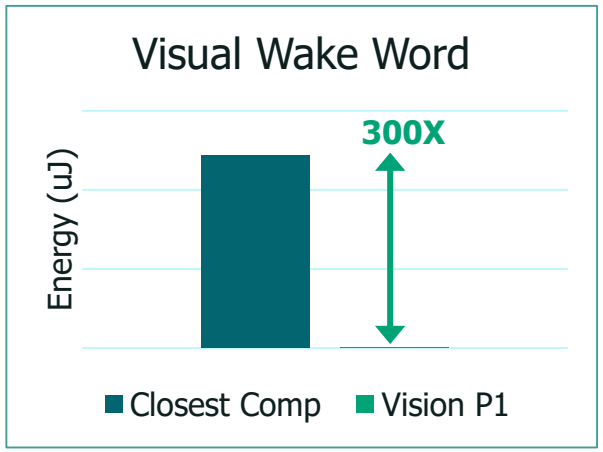


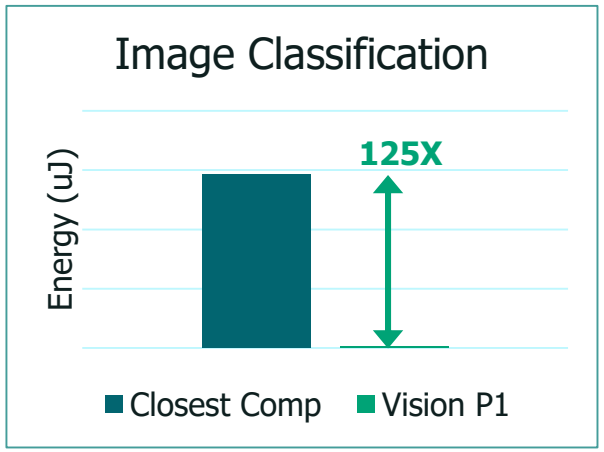
TinyML v0.5 Benchmark Comparison

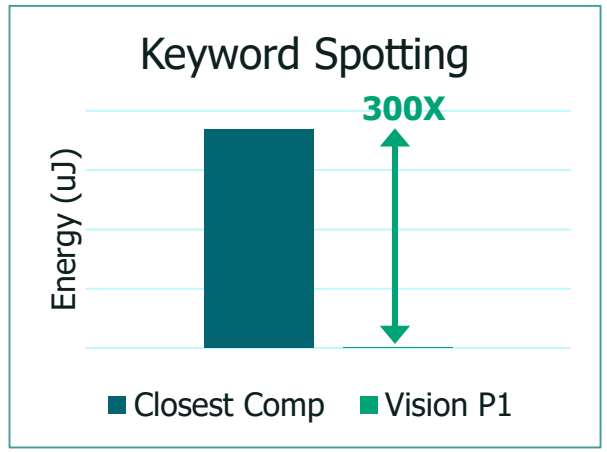


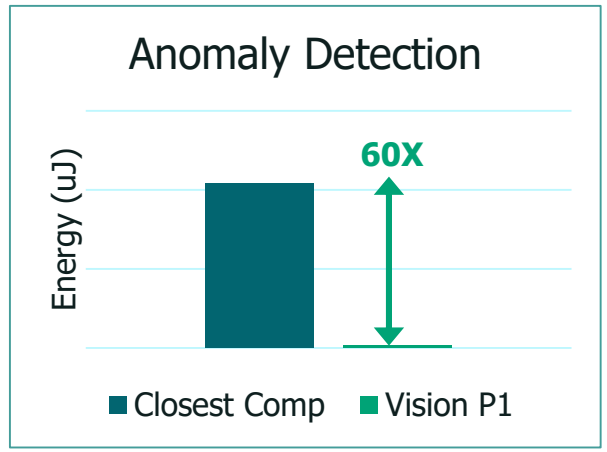












- ✓ Tensilica® Vision P1 DSP processes vision and audio workloads for always-on
- ✓ Energy efficient and customizable to run at different frequencies
- ✓ <1 mW always-on solution

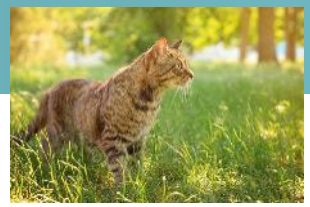
RTL simulated on P&R netlist

Source: https://mlcommons.org/en/inference-tiny-05/

Other Vision Capabilities

Smart Image Sensor

Tensilica®
Vision P1 DSP



Normal Image
(Pre/Post ISP)

Smart Sensor Processing

Region of
Interest Crop

Cat

Metadata

Bounding Box

Segmentation

Foveated Rendering

Headset with pupil camera

Cameras capture
pupil images

Tensilica
Vision P1 DSP

Detects
and gaz

Gaze direction used for
foveated rendering

- ✓ Supports all imaging and vision DSP tools and libraries
- ✓ Supports AI flow of XNNC, NNAPI, TFLm, and more
- ✓ 100+ model zoo networks supported out of the box

cadence

One Last Thing…

Come visit Cadence booth located at 317

Sources:

MLCommon Tiny Networks Performance
https://mlcommons.org/en/inference-tiny-05/

Always-On Market Trend
https://www.grandviewresearch.com/industry-analysis/light-sensor-market

2022 Embedded Vision Summit

- **Tensilica® Vision P1 DSP Always-On FPGA Demo**
  - Visual Wake Word + Keyword Spotting Demo on same IP
- **Other Vision and AI Demos**